Maschinelle Sprachverarbeitung

Exercises, Assignment 1: Parsing an XML corpus

Ulf Leser

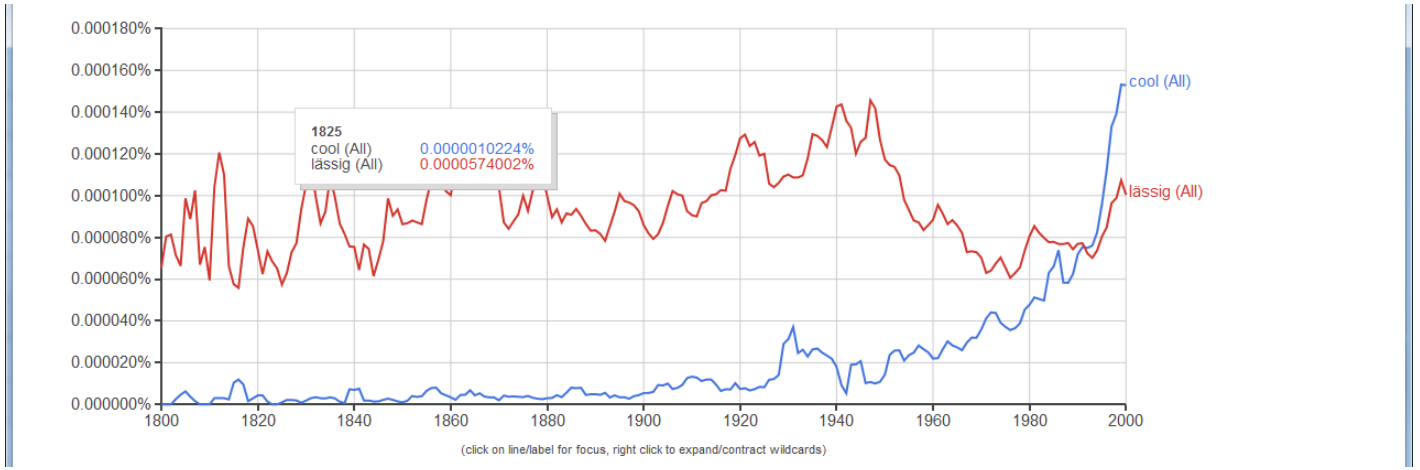
Assignment

- Download archive of the Reuters corpus
  - modnlp.berlios.de/reuters21578.html
- Parse the XML files (reut2-000 to reut2-021; excluding reut2-017)
  - Directly use a XML parser (SAX or DOM or write your own one)
  - No external libraries allowed (e.g. Saxon)
  - No need to do validation, you may ignore the DTD

Properties to Count

- Total number of documents
- Total number of token in TEXT/TITLE and TEXT/BODY
  - Use whitespace (\s) tokenization
  - Convert all token to lower case
- Number of entities of the following types
  - Topics (Distinct and total)
  - Places (Distinct and total)
  - People (Distinct and total)
- Overall frequency for all tokens in TITLE / BODY
  - Return token and frequency of top-100 token
  - Plot distribution of all token
  - What kind of distribution can be observed?

Google N-gram viewer



Details

- Use Java (or JVM compatible: Scala, Jython, …)
- Submit one JAR file called groupX.jar
- Must start with „java -jar groupX.jar directory
  - Parse all and only files in „directory“
  - These will all be corpus files
- Include source code
- Test your JAR before submission
- PDF des Plots
- Output to stdout
- Send solutions by 8.11.2015 latest
  - Mail to Ulf Leser
  - This is an unusual long period

Competition

- Implement fastest (and correct!) solution
- We measure wall clock time on GRUENAU4 using “time“
  - Parallelization makes sense
  - We cannot assure full availability of the machine